RESEARCH PLAN PROPOSAL

DESIGN AND IMPLEMENTATION OF A NOVEL PREPROCESSING ALGORITHM TO DISCOVER USER NAVIGATION PATTERNS IN THE WEB

For registration to the degree of

Doctor of Philosophy

IN THE FACULTY OF SCIENCE

THE IIS UNIVERSITY, JAIPUR

Submitted by

Priyanka Verma

Enroll. No. ICG/2013/16382

Under the Supervision of

Dr. Nishtha Kesswani

Assistant Professor

Department of Computer Science and IT

May 2014

Table of Contents

1. TOPIC 3

2. INTRODUCTION 3

2.1 Web Mining 3

2.2 Applications of Web Mining 3

2.3 Types of Web Mining 3

2.4 What is Web Usage Mining 3

2.5 Stages of Web Usage Mining 4

2.6 Applications of Web Usage Mining 4

2.7 Preprocessing stage of web Usage mining 4

2.7.1 Log Files 5

2.7.2 Basic Access Log Format 5

2.7.3 Preprocessing stages and their algorithms 5

3. REVIEW OF LITERATURE 8

4. MOTIVATION, JUSTIFICATION AND RELEVANCE 9

5. OBJECTIVES 9

6. PLAN OF WORK 10

6.1 Research Design 10

6.2 Methodology 10

6.3 Work Plan 11

7. PLACE OF WORK 12

8. LIMITATIONS AND ALTERNATIVE PLAN OF STUDY 12

9. REFERENCES 12

1. TOPIC

Design and Implementation of a Novel Preprocessing Algorithm to Discover User Navigation Patterns in the Web

2. INTRODUCTION

2.1 Web Mining

Web site is a group of web pages placed under the some domain name. These web sites have to be accessed through domain under which they are created. Web pages may contain text, images, videos, hyperlinks through which navigation happens. Whenever user accesses any website, log files are created. Log files record information about each user's website access. Due to increase in the usage of web the size of log files is increasing day by day. Data in Web log files is stored in various formats such as the NCSAs Common log file format, the W3C Extended Log File format or the IIS log file format. Log files are located in different locations like web server, proxy server, and Client browser. For getting optimum results we need to extract data from all three log files. The ability to know the patterns of user's habits and interests helps in increasing performance of web site by improving web site design. Web mining is the application of various data mining techniques to retrieve, extract and evaluate information for knowledge discovery from web documents and services.

2.2 Applications of Web Mining

There are various applications of web mining as given by [1]

- Personalized experience

Web mining can be useful in Amazon.com, Yahoo.com to analyze user's past behavior to send personalized messages and also to set up special offers, recommend products based on user's location.

- Google web search

Web mining is useful for content analysis to determine pages relevant to topic and also for ranking pages based on their quality so that whenever the user enters search query highest rank pages relevant to search query would be returned by the search engine

- Auction sites

Web mining is helpful in various auction sites also such as e Bay to categorize participants into various types also it would help to determine fraudulent bids.

2.3 Types of Web Mining

Web mining tasks can be categorized into three types

- Web Structure Mining: Web structure mining discovers useful knowledge from the structure of web site i.e. from hyperlinks which link the entire site. For example, from the links, we can discover important Web pages, which is technology used in search engines. We can also discover users who share common interests.
- Web Content Mining: Web content mining extracts or mines useful information or knowledge from the contents of the web page . For example, we can automatically classify and cluster Web pages according to their topics.
- Web Usage Mining: Web usage mining is the discovery of access patterns followed by the user, from Web logs that record all the activities of the user.

2.4 What is Web Usage Mining

Web usage mining (WUM) or web log mining, is revealing user's behavior or interests by applying various data mining techniques on log files stored on server, client or proxy. Knowing the patterns of user's habits and interests helps in efficient building of various applications. The main source of data for web usage mining

consists of various logs stored on numerous web servers all around the world. WUM is also known as Clickstream Analysis. A clickstream is the aggregate of sequence of page visits by a particular user navigating through a Web site.

2.5 Stages of Web Usage Mining

There are four stages in web usage mining.

- Data Collection: During this phase users log data is collected from various sources like server side, client side, proxy servers and so on.
- Preprocessing: During this phase a series of steps of processing of log files are performed which includes Data Cleaning, User Identification, Session Identification and Path Completion.
- Pattern discovery: This phase is application of various data mining techniques like association, clustering to processed data obtained after preprocessing phase to discover various patterns.
- Pattern analysis: After patterns are discovered from web logs analysis is done using various query mechanism such as SQL to perform various OLAP operations

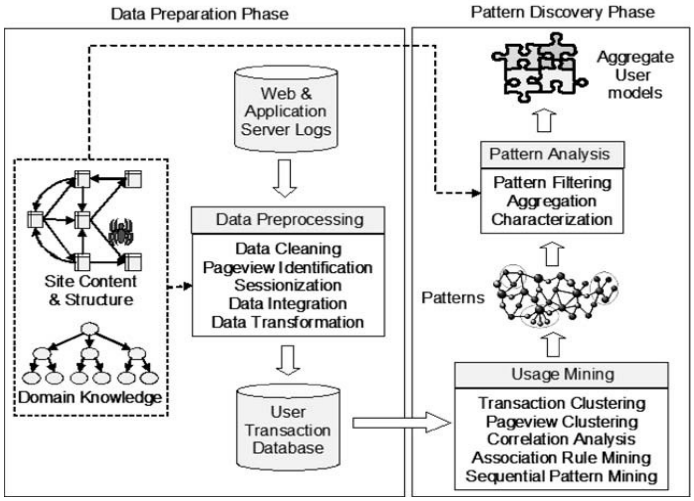


Fig. 1. Stages of web usage mining [2]

2.6 Applications of Web Usage Mining

There are various applications of Web Usage Mining as given by [1]

- Web usage mining is useful in various Ecommerce applications for
  - Determining lifetime value of clients
  - Designing cross marketing strategies across products
  - Evaluating various promotional campaigns
  - Targeting electronic ads and coupons at user groups based on their access patterns
  - Predicting user's behavior based on previously learned user's profile
  - Presenting personalized information to users based on their interests and profiles
- Web usage mining is also useful in efficient web site management by
  - Determining the best way to structure the Web site
  - Identifying "weak links" for elimination
  - Pre-fetching files that are most likely to be accessed

2.7 Preprocessing stage of web Usage mining

When loading a particular web page, the browser also requests for all the objects embedded in that page, such as .gif or .jpg graphics files. Each such request is logged separately including requests for .gif or .jpeg files. Thus

click stream data requires substantial preprocessing before the user's behavior can be analyzed. Preprocessing is opening and most essential step. Input to the preprocessing phase of Web usage mining is Web Log file. It is large in size and contains number of raw and irrelevant entries and therefore cannot be directly used in Web usage mining process. Preprocessing of log fie is a very complex and painstaking job and it takes 80% of the total time of web usage mining process as whole.

2.7.1 Log Files

Log files are plain text files created by servers and it contains information about user's requests made on a particular Web site. For each request from a user's browser to a web server, a response is generated automatically, called a web log file or log file, or web log. Log files can either be Access logs, Agent logs, Error logs or Referrer logs[10].We would focus on access logs as we have analyze access patterns followed by the user. Access log file may either be comma-delimited, space-delimited, or tab-delimited. These Log file gives information about:

1. Pages requested in website?
2. Bytes sent to user from server?
3. Type of error that occurred?

Web log file can be of three different types depending upon location.

- Web server logs
- Web proxy server
- Client browser:

A sample web log is shown below.

```
01-12-2011 12:36:22 172.186.210.86 - CSLNTSVR1 - 192.168.12.44 80
GET /web-inf/img/icons/img1.gif 304 140 470 0 HTTP/1.1
www.onlinelearning.com.my GoogleChrome/2.0+compatible;+MSIE+5.5;
+Windows+7;
+Win+9x+4.90ASPSESSIONIDCSTSBQDC=NBKBCPIBBJCMMFIKMN
N KFD;+browser=done; +ASPSESSIONIDAQRRCQCC=LB
DGBPIBDFCOKHMLHEHNKFBN www.onlinelearning.com.my
```

Fig. 2. Web Log File

2.7.2 Basic Access Log Format

Access log files contains information as given by [10]

- Date - The format is DD-MM-YYYY
- Time -The time format is HH:MM:SS
- Client IP Address
- Server Name
- Server IP Address
- Server Port
- Server Method
- Server URI-Query
- Status -Success (200 series), Redirect (300 series), Failure (400 series) and Server Error (500 series).
- Transfer Volume (Bytes) Field
- Referrer Field
- User Agent Field

2.7.3 Preprocessing stages and their algorithms

- **Field Extraction from Log Files and Storage in Table**

Extracting the log files and storing them in the table is the first step in preprocessing. The log entry contains various fields which have to be separated for the processing. The process of separating fields from the log file is known as Field Extraction. The server uses different characters as separators for fields. The most commonly used separator character is ',' or 'space' character. Step by step algorithm for reading server logs and storing them in the table is given by [10] [16].

1 Create procedure read for reading log file
2 Set record counter variable as zero
3 Create variable log reader that will create server object
4 Through log reader open log file specified in read procedure
5 Read file till the end of file is reached
6 At each record increment record counter to keep track of number of records in log file
7 Open DB connection
8 Create a table to store log data (LT)
9 Extract all fields stored in log file created above and add into the Log Table (LT)
10 Close DB connection and Log File

- **Data Cleaning**

Data cleaning is second stage after the storage of log file in table. This stage is performed to remove the unnecessary content which includes requests for images, styles and scripts or other files. This stage is the most simplest of all stages as it consists of only filtration of the data. During this stage requests made by the crawlers are also deleted .Final file after cleaning contains accesses made by standard visitors only [10]. Step by step algorithm for Data Cleaning is given by [16]

1. Read records in LT (Log Table)
2. For each record in LT
3. Read fields (Status code, requestmethod)
4. If Status code='200'and requestmethod= 'GET'
   Then
5. Get IP_address and URL_link
6. If suffix.URL_Link= {*.gif,*.jpg,*.css}
   Then
7. Remove suffix.URL_link
8. Save IP_sddress and URL_Link
   End if
9. Else Next record
   End if
   End

- **User Identification**

This step identifies individual users by using their IP address. "If IP address is same but browser version or operating system is different then it represents different user." [2] .Step by step algorithm for User Identification is given by [17]

Begin
1. Read records in LT
2. For each record in LT
3. Read fields (IP Address, User Agent)
4. If IP Address of first log entry = IP Address of second log entry.
   Then
5. If User Agent of first log entry = User Agent of second log entry
   Then
6. User is same
   End if
   Else
7. User is different
   End if
8. Next record
   End

- **Session Identification**

Session is a sequence of requests made by a single user with a unique IP address on a Web site during a specified period of time. Sessions are identified either through time based heuristics or navigation based heuristics. "Time oriented heuristics is based on total session time or page-stay time. Navigation oriented approach makes use of uses the web graph constructed by using hyperlinks among web pages".[15] [18] "Smart SRA stands for Smart Session construction Algorithm which is designed to overcome deficiencies of

the time and the navigation oriented heuristics. Smart-SRA uses a novel approach to construct user session as a set of paths in the web graph where each path corresponds to user's navigations among web pages. That is, in Smart-SRA, server request log sequences are processed to reconstruct web user session not as a sequence of page requests, but, as a set of valid navigation paths. The two main phases of Smart-SRA are explained below:

- In the first phase, the access data stream of web users are partitioned into shorter page request sequences called candidate sessions, by using session duration time and page-stay time rules.
- In the second phase, candidate sessions are divided into maximal sub-sessions such that for each consecutive page pair in the sequence there exists a link from previous one to latter one. At the same time, page stay time rule for consecutive pages is also satisfied."[11]

In other approach given by GuerBas et al. [3] the original time oriented sessionistic algorithm is modified to make sure that identified sessions from the same sequence share a pattern i.e. there should be some specified pages in common between sessions. Algorithm for same is given below

```
Input: a list of sequences
Output: a list S of valid sessions
begin
01 - S ← {}
02 - Sort sequences according to user id and timestamp
03 - For all sequences do
04 -        Divide the current sequence according to page-stay threshold
05 -        If the sequence is monolithic
06 -            current sequence is a single session
07 -            S ←S ∪ {current sequence}
08 -        Else /*sequence with gaps*/
09 -            If there is a shared pattern between the sub-sequences
10 -                Each subsequence is a separate session and all sessions are from one single user
11-                     S ←S ∪ {All su-bsequences}
12 -        Else /* no shared pattern*/
13-             Skip sequence.
14- End for
15- Return
End
```

Fig. 3.Algorithm for session identification [3]

- **Path Completion**

"Path completion is a process of adding the page accesses that are not in the web log but those which have actually occurred. Client or proxy side caching can often result in missing access references to those pages or objects that have been cached. For instance, if a user returns to a page A during the same session, the second access to page A will likely result in viewing the previously downloaded version of page A that was cached on the client-side, and therefore, no request is made to the server. This results in the second reference to page A not being recorded on the server logs."[2]

- **Data Integration**

In order to provide the most effectual framework for pattern discovery, data from a variety of other sources has to be integrated with the preprocessed clickstream data. This is particularly relevant in case in e-commerce applications where the integration of both user data and product attributes and categories from operational databases in the mining process can allow for the discovery of important metrics such as customer adaptation ratios and lifetime values.[2]

We can conclude that preprocessing is an important stage and it takes 80% of the effort of web usage mining process and this stage if done correctly will make the process of pattern identification easier. Our main focus would be on studying existing preprocessing algorithms, finding out errors in those algorithms and propose a new preprocessing algorithm.

3. REVIEW OF LITERATURE

Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from Web data, specifically web logs, in order to perform improvements in web based applications. Web usage mining consists of three phases, pre-processing, pattern discovery, and pattern analysis. After the completion of these three phases the user can find the required usage patterns and use this information for specific needs. The goal of Web Usage Mining is to extract the useful information from web data or web log files. The other goals are to enhance the usability of the web information and to apply the technology on the web applications for pre-fetching and catching, personalization etc. For decision management, the result of web usage mining can be used for target advertisement, improving web design, improving satisfaction of customer, guiding the strategy decision of the enterprise and market analysis [12].

To track any transaction made on the web we use log files. Log files can either be Access log files, Referrer log files, Agent log files or Error log files. We would be focusing on Access log files. Access log files store the data about all access to pages made by the user. It consists of various attributes which have already been mentioned (Refer 2.7.2). "The Web logs are often available in two formats:CLF(Common Log Format) and ECLF (Extended Common Log format)"[7]

In order to get the suitable Web log data to perform the mining to discover useful patterns, we must undertake a series of operations (Preprocessing) on the original Web log files such as data cleaning, user and transaction identification, data integration and so on. [7]. Data cleaning is also known as data cleanup. Data cleaning is a process in which data errors, inconsistencies are removed. The data cleaning process removes the noise or irrelevant data which includes jpeg ,gif files or script images. The size of log file, after successful completion of the data cleaning step is reduced [10].Next step of preprocessing is user identification, For user identification firstly, if the IP addresses are identical, but the browser software or the operating systems are different in Agent information, then two different users can be assumed. Secondly, if IP address and the Agent information are identical, then we should judge whether there is a connection between the pages which are requested to access and the pages which have been accessed. If there is no direct link between them, then we can assume that there exist multiple users in the machine which accesses the Web sites [7]. Next stage is Session identification. Session is a sequence of requests made by a single user with a unique IP address on a Web site during a specified period of time. There are two approaches for session identification. First approach is Smart-SRA(Smart Session construction) Algorithm which is designed to overcome deficiencies of the time and the navigation oriented heuristics. Smart-SRA uses a novel approach to construct user session as a set of paths in the web graph where each path corresponds to users' navigations among web pages [11]. In other approach given by GuerBas et al.[3] the original time oriented sessionistic algorithm is modified to make sure that identified sessions from the same sequence share a pattern i.e. there should be some specified pages in common between sessions [3]. Next stage is Path completion which is a process of adding the page accesses that are not in the web log but those which have actually occurred. Missing page references occur due to caching. [2].

Till date there are lot of tools and systems proposed to implement web usage mining .One such tool is 'A Pre-processing Tool for Web Usage Mining in the Distance Education Domain'. This tool extracts access patterns that reveal students behavior which can be further analyzed to improve learning process[6].Another system that has been proposed is 'SUGGEST : AWeb Usage Mining System'. This system is designed to integrate the WUM process with web server functionalities to provide useful information which makes navigation process easier for the user and improves the server performance. It mainly extracts useful information from users access data which can then be used to personalize, give suggestions to user. [5]

The preprocessing stage of web usage mining has been implemented through various platforms making use of algorithms given above[Refer 2.7.3].One such implementation is done by Roop Ranjan[12].He proposed a framework for web mining on log files of dynamic web site. He has used Log Parser to parse log files. Log parser parses Log files which are large in size, For example to find all requests made by particular IP address. He has

used .Net Framework and SQL server. After parsing log files the log files are read and are displayed on web page in desired format. Then log data can be analyzed and reports can be generated.

Web usage mining has various applications such as it can be used to mine 'code search engine' log to observe usage behavior of user's, how user's express their search term in query and to find out topics searched so that we can know what developers usually look for, so that site content can be better organized [4].WUM can also be used to enhance search results i.e. to help users to find relevant content by recommending related news articles and blog posts. Graph based approach is used to quickly search and recommend news articles to user.[8].Another application of WUM is periodic Web Personalization which aims to recommend the most relevant resources to the user during specific time period by analyzing the periodic access patterns of the user from the web usage logs. It personalizes content and structure of web site according to needs of user by analyzing information stored in logs. With periodic personalization we can find out resources user is interested in during a specific time period [9].Web usage mining can also be used to analyze data that comes from educational system to find most effective structure of course, optimize learning content, recommend the most suitable learning path based on students behavior and provide more personalized content.[14]

4. MOTIVATION, JUSTIFICATION AND RELEVANCE

The main motivation behind choosing this topic is that the information stored in World Wide Web is increasing day by day and so is the user's demand to get right data. For getting right data user's navigational patterns have to be studied. For this various mining techniques have to be used which will help us to extract meaningful patterns and relationship from large data [6].In the field of web mining researchers have done considerable research but rapid development of internet makes these studies lag behind [7] Web mining can be divided into three main domains out of which our main focus is on Web Usage Mining [8].The importance of Web Usage Mining is increasing day by day to discover the access patterns followed by the user to personalize web sites, target various campaigns at user etc. Amongst all the stages of web usage mining preprocessing is considered to be the lengthiest step and it takes 80% of all the effort and this step if performed right can make pattern discovery and analysis easy. A web log file stores user's activities performed during a given period of time. They offer valuable information about the various resources accessed on the web site. By analyzing the patterns stored in these log files we can improve the content of web site. Lot of work on proposing algorithms for various preprocessing stages is already ongoing. But as already mentioned with the advancement in web there are a lot of new concepts which have to be added to these preprocessing algorithms. Our main focus would be on finding out errors/problems in existing algorithms and propose a new preprocessing algorithm.

5. OBJECTIVES

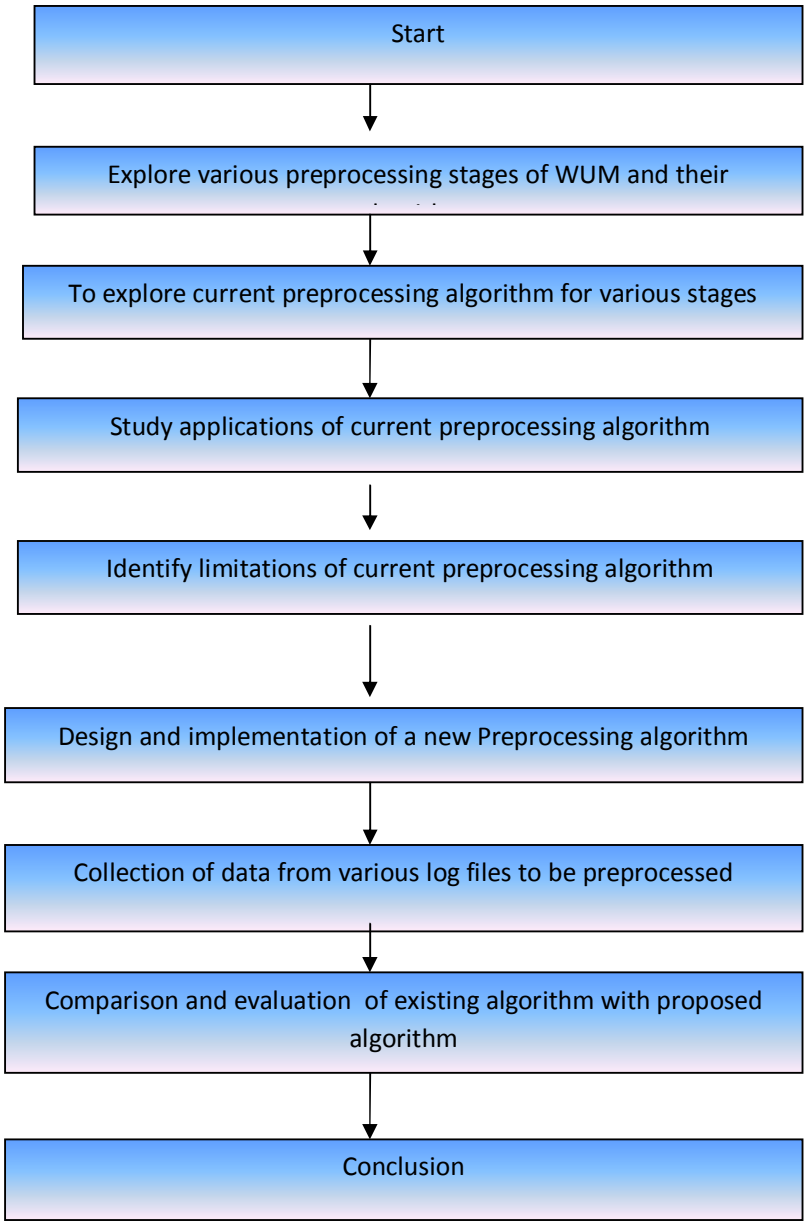
- Detailed study of Web Usage Mining techniques
- To explore various preprocessing stages of Web Usage Mining and algorithms
- To explore current preprocessing algorithm for various stages
- To study application of current preprocessing algorithm to identify navigation patterns followed by the user
- To identify the limitations of current preprocessing algorithms
- To design and implement new preprocessing algorithms
- To compare and evaluate results of proposed algorithms with existing algorithms

6. PLAN OF WORK

6.1 Research Design

Logs from various sources would be collected and would be preprocessed through preprocessing algorithm. The framework for implementing preprocessing algorithms would be developed and tested.

6.2 Methodology


Start
Explore various preprocessing stages of WUM and their
To explore current preprocessing algorithm for various stages
Study applications of current preprocessing algorithm
Identify limitations of current preprocessing algorithm
Design and implementation of a new Preprocessing algorithm
Collection of data from various log files to be preprocessed
Comparison and evaluation of existing algorithm with proposed algorithm
Conclusion

6.3 Work Plan

Phase 1: To study current approaches for preprocessing web usage data

Phase 2: Study current preprocessing algorithms and their application to discover user navigational patterns

Phase 3: Investigate problems with existing algorithms

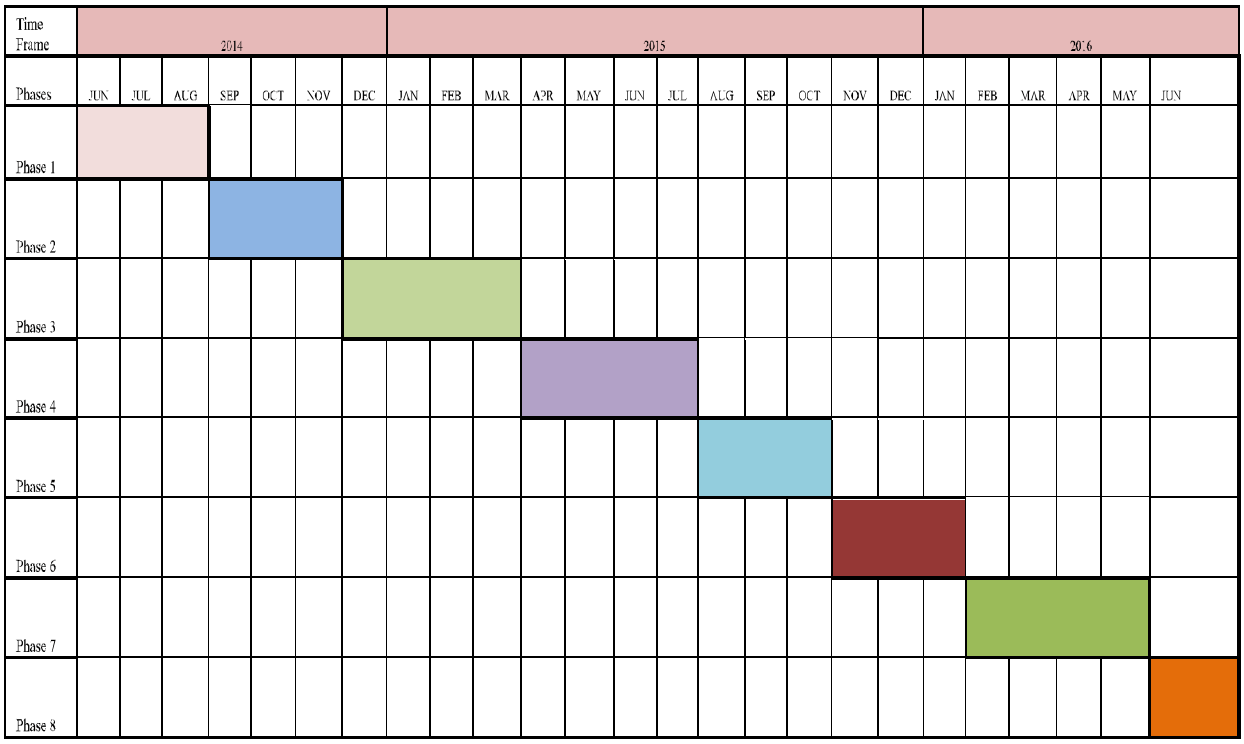
Phase 4: Propose a new Preprocessing algorithm

Phase 5: Collection of data from various log files to be preprocessed

Phase 6: Compare results of existing algorithm with new proposed algorithm on data collected from various log files

Phase 7: Result Generation and Thesis Writing

Phase 8: Thesis submission



7. PLACE OF WORK

The entire setup for running preprocessing algorithms, both proposed and existing would be set up in Information Technology lab in The IIS University campus.

8. LIMITATIONS AND ALTERNATIVE PLAN OF STUDY

The data collection from various logs is required for the research to be conducted. This data will be have to be arranged either from university campus or from other external sources.

9. REFERENCES

[1] Jaideep Srivastava, "Mining :Accomplishments & Future Directions",USA,ch.3,pp. 51-64,2003

[2] Bamshad Mobasher, "Web Usage Mining" in Web DataMining Exploring Hyperlinks,Contents, and Usage Data,Chicago,USA,Springer,ch.12,pp. 449-482,2012

[3]Guerbas et al., "Effective web log mining and online navigational pattern prediction", Knowledge-Based Systems,Elsevier,vol. 49,pp. 50-62, 2013

[4] Sushil Krishna Bajracharya and Cristina Videira Lopes, "Analyzing and mining a code search engine usage log" in Empirical Software Engineering,Springer,vol.17, doi:[DOI 10.1007/s10664-010-9144-6] ,pp. 424-446,2012

[5] Ranieri Baraglia, Paolo Palmerini, " SUGGEST : AWeb Usage Mining System" in Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC.02),IEEE,pp. 282-287,2002

[6] Carlos G. Marquardt Karin Becker Duncan D. Ruiz, "A Pre-processing Tool for Web Usage Mining in the Distance Education Domain" in Proceedings of the International Database Engineering and Applications Symposium (IDEAS'04),IEEE,pp. 78-87,2004

[7] Ma Shu-yue, Liu Wen-cai Wang Shuo, "The Study on the Preprocessing in Web Log Mining" in proceedings of the Knowledge acquisition and modelling(KAM),IEEE,pp. 315-317,2011

[8] Ida Mele, "Web Usage Mining for Enhancing Search-Result Delivery and Helping Users to Find Interesting Web Content" in proceedings of WSDM'13,ACM,pp.765-769,2013

[9] Baoyao Zhou1, Siu Cheung Hui and Alvis C. M. Fong, "An Effective Approach for Periodic Web Personalization"in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), pp.284-292,2006

[10] K Sudheer Reddy, Dr. G. Partha Saradhi Varma, Dr. I. Ramesh Babu, "Preprocessing the Web Server Logs An illustrative approach for effective usage mining" in ACM SIGSOFT Software Engineering Notes Page,vol.37, doi:[http://doi.acm.org/10.1145/180921.2180940], pp. 1-5,2012

[11] Murat Ali Bayir, Ismail Hakki Toroslu, "Smart Miner: A New Framework for Mining Large Scale Web Usage Data" in proceedings of International World Wide Web Conference Committee (IW3C2),ACM ,pp. 161-170,2009

[12] Roop Ranjan, Sameena Naaz, Neeraj Kaushik, "Web Miner: A Tool for Discovery of Usage Patterns From Web Data", International Journal on Computer Science and Engineering (IJCSE), vol. 5,pp. 286-293,2013

[13] Michal Munka, Jozef Kapustaa, Peter Švec, "Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor" in proceedings of International Conference on Computational Science(ICCS 2010)[Online] Available:www.sciencedirect.com,doi:[doi:10.1016/j.procs.2010.04.255], Procedia Computer Science ,vol.1,Elsevier,pp. 2273-2280,2010

[14] Michal Munka, Martin Drlík, "Impact of Different Pre-Processing Tasks on Effective Identification of Users' Behavioral Patterns in Web-based Educational System" in proceedings of International Conference on Computational Science, ICCS 2011[Online] Available:www.sciencedirect.com, doi:[http://dx.doi.org/10.1016/j.procs.2011.04.177],Procedia Computer Science, vol.4,Elsevier,pp. 1640–1649,2011

[15] V.Chitraa , Dr. Antony Selvdoss Davamani,"A Survey on Preprocessing Methods for Web Usage Data", International Journal of Computer Science and Information Security(IJCSIS),vol. 7,pp.78-83,2010

[16] Priyanka Patil, Ujwala Patil, "Preprocessing of web server log file for web mining" in proceedings of National Conference on Emerging Trends in Computer Technology (NCETCT 2012) [Online]Availaible:www. worldjournalofscience.com,World Journal of Science and Technology,vol. 2, pp. 14-18,2012

[17] V.Chitraa ,Dr.Antony Selvadoss Thanamani, "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", International Journal of Computer Applications,vol. 34, pp. 23-27,2011

[18] R. Cooley, B. Mobasher, J. Srivastava. ,"Data preparation for mining world wide web browsing patterns." ,Knowl. Inf. Syst., Springer,vol. 1,pp. 1–27, 1999.

[19] R. Cooley, P.-N. Tan, and J. Srivastava.," Discovery of interesting usage patterns from web data". ,WEBKDD, pp. 163–182, 1999.

[20] D. Donato, L. Laura, S. Leonardi, and S. Millozzi., "The web as a graph: How far we are.", ACM Transactions on Internet Technology (TOIT),vol. 7 , pp. 1-25,2007.

[21] R. Kumar et al., "The web as a graph." in Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems,ACM, pp. 1–10, 2000.

[22] E. Frias-Martinez and V. Karamcheti., " A customizable behavior model for temporal prediction of web user sequences.",WEBKDD, pp. 66–85, 2002

[23] N. M. Abo El-Yazeed, "An overview of preprocessing of web log files for web usage mining",[Online] Availaible : etms-eg.org/cms/upload/1354650330.pdf, pp. 1-16.